
ISSN: 2321-2152
IJMECE
International Journal of modern
electronics and communication engineering
E-Mail
editor.ijmece@gmail.com
editor@ijmece.com
www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

Hypothyroidism Early Detection Prediction by Feature Selection and Classification Methods

[1]DR. P. Ramasubramanian, [2] T.Sathwika,

[1] Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract

When it comes to females in Bangladesh, thyroid illness ranks high. One of the most prevalent forms of thyroid illness is hypothyroidism. It is rather evident that the majority of people diagnosed with hypothyroidism are female. The sickness is quickly becoming a serious illness since most people are unaware of it. Preventing it from progressing to a more dangerous level requires early detection so that physicians can provide more effective medicine. Machine learning illness prediction is challenging. In order to make accurate illness predictions, machine learning is crucial. The use of separate feature selection methods has once again made illness prediction and assumption easier. The thyroid may develop into either hyperthyroidism or hypothyroidism. Attempting to forecast hypothyroidism in its early stages is the focus of this article. We have accomplished this primarily via the use of three feature selection strategies in conjunction with several categorization algorithms. In addition to classification algorithms such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB), we employ feature selection techniques such as Recursive Feature Selection (RFE), Univariate Feature Selection. It is reasonable to conclude from the data that all four of the classification algorithms benefit from the consistent 99.35% accuracy we get using the RFE feature selection method. Therefore, our study concludes that RFE improves the accuracy of each classifier compared to all other feature selection approaches.

Keywords

Classification, Machine learning, Recursive Feature Selection, Data mining, Thyroid illness

I. INTRODUCTION

At this time, thyroid illness is among the most serious health problems that people face, and it may soon become epidemic among women. Experts estimate that 50 million individuals in Bangladesh are affected by thyroid illness. Thyroid illness affects women at a rate ten times higher than males. Even though half a million individuals have thyroid illness, about 30million of them people have no idea that they have it. Between twenty and thirty percent of women have thyroid illness, according to research from the Bangladesh Endocrine Society (BES) [14]. at humans, the thyroid gland is located at the exact center of the neck. It has a little size and a butterfly form. To regulate different bodily functions, it secretes a cocktail of hormones that are transported throughout the body in the blood. Thyroid hormone regulates metabolism, sleep, development, libido, and mood. Feelings of lethargy, restlessness, and even weight loss are all linked to fluctuations in thyroid hormone release. Thyroxin (T4) and Triiodothyronine (T3) are the two primary thyroid hormones. Keeping our energy levels stable is mostly the job of these two hormones. The pituitary gland secretes Thyroid Stimulating Hormone (TSH), which the thyroid gland uses to secrete T3 and T4. Among thyroid disorders, hypothyroidism is the most prevalent. 2) Overactive thyroid. Hypothyroidism occurs when the thyroid gland is unable to produce an adequate amount of thyroid hormones, leading to elevated TSH levels and decreased T3 and T4 levels. Loss of appetite, lethargy, mental fog, etc. are some of the symptoms it manifests. Hyperthyroidism occurs when the thyroid gland generates an excess of thyroid hormone (THC) relative to what the body requires, leading to elevated T3 and T4 levels and decreased TSH. Hair loss, nervousness, excessive perspiration, and other symptoms may be present. Since hypothyroidism is the most prevalent among Bangladeshi women, we have focused on this condition in our studies. Consequently, the first stage of hypothyroidism detection was the major focus of our investigation. Machine learning has recently exploded in popularity as a tool for illness detection. The use of machine learning algorithms for illness presumption is both practical and efficient. In this work, we have used feature selection and classification methods for primary stage

hypothyroidism prediction. A licensed diagnostic facility in Dhaka, Bangladesh was the source of the data we used. All things considered, we have amassed a sizable amount of data including nine distinct qualities. Of these records, 77% are to females and the remaining 4% to men. A variety of classification algorithms, including Support Vector Machine (SVM), Decision Tree, Logistic Regression (LR), Random Forest (RF), and Naive Bayes (NB), as well as three feature selection techniques—Recursive Feature Elimination (RFE), Univariate Feature Selection (UFS), and Principal Component Analysis (PCA)—are fundamental to our work. After much trial and error, we found that RFE feature selection improves accuracy regardless of the classification method.

II. LITERATURE REVIEW

Using the feature selection strategy and classification for hypothyroidism prediction, your approach essentially suggests a model to identify hypothyroidism in its early stages. A number of relevant approaches have emerged in recent years, and we'll go over a few of them here. The authors of [1] suggested a method for early detection of cardiac disease using classification and regression trees. The long-term objective of this planned project is to develop a system for diagnosing heart problems; this will help cut down on the number of needless echocardiograms and save babies born with heart defects from being released into the world. Additionally, they use Classification & Regression Tree (CART) to examine PCG (phonocardiogram) data in this study. They use time and frequency feature extraction in categorization. Additionally, k-means clustering is used. One way to build a CART regression tree, which is a kind of binary decision tree, is to divide each node into two daughter nodes. On the experimental dataset, they achieved a specificity of 98.28%, a sensitivity of 100%, and an accuracy of 99.14%. An Intelligent System for Classification and Diagnosis of Thyroid Diseases [2] was the subject of this investigation. In order to classify and diagnose thyroid illness, they suggested using Weighted SVM classification and particle swarm optimization to improve SVM parameters like TSH, T3, and T4. This would allow for early detection of the ailment. In addition, they take user input and utilize KNN to approximatively determine the missing value. Using data mining techniques, the authors of [3] built a system to predict thyroid disease. It suggests a way to use a classification algorithm like KNN, SVM, Naive Bayes, or decision tree C4.5 & ID3 algorithm to discover a way to detect thyroid illness at an early stage with greater accuracy. A study that aimed to improve the diagnosis of thyroid disease using feature selection algorithms was suggested in [4]. The goal of this study is to examine the impact of using filter-based feature selection algorithms (F-Score) and wrapper-based techniques (Recursive Feature Elimination) on illness identification and classification. In addition, four classifiers—Extreme Learning Machine, Multilayer Perceptron, Back Propagation Neural Network, and Support Vector Machine—were used. The wrapper-based approach achieved the highest possible efficiency and accuracy of 98.14 percent when used with an ELM classifier. In their study, the authors suggested utilizing PNN and SVM to build a genetic algorithm-based technique for diagnosing thyroid disease [5]. In order to distinguish between hypothyroid and hyperthyroid cases in diagnostics, this framework suggested using Support Vector Machines (SVM) and Probabilistic Neural Networks (PNN) for classification. They used a genetic algorithm for feature selection. With SVM&PNN and GA (FS), their accuracy is 100%. The authors provide an enhanced ensemble classification method for thyroid disease using random forest in their work [6]. A novel random forest-based approach for thyroid illness categorization was suggested in this study. They reached 96.16 percent accuracy using a random forest-based ensemble classifier approach. An Interactive System for the Prediction of Thyroid Disease Using Machine Learning Techniques [7]. Classification has been carried out using the dataset from the UCI repository. To evaluate the likelihood of thyroid illness in patients, the authors of this suggested study employed Machine Learning Algorithms such SVM(99.63), K-NN(98.62), Decision Trees(75.76), and ANN (97.5). Ranker Search and Naive Bayes, two feature selection and classifier algorithms, provide an accuracy of 95.38% in this study's survey of thyroid problem diagnoses [8]. Using an Optimized SVM Method, this study built Hypothyroid Disorder Classification [9]. This study presented a strategy for hypothyroid disease level detection utilizing a combination of artificial neural networks (ANNs), logistic regression (LR), K-nearest neighbor (KNN), and support vector machines (SVM). Among the four classifiers tested, the Logistic Regression approach had the best accuracy at 96.08%; however, following data standardization and parameter adjustment, SVM had the maximum accuracy at 99.08%. The authors of this study suggested a technique for classifying thyroid data using a kernel-based classifier process and optimum feature selection in their publication

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

[10]. Using enhanced gray wolf optimization, this suggested model's originality and purpose are to increase the performance of the classifying process via feature selection. This method use MKSVM to differentiate between thyroid illnesses with an impressive 98.65% accuracy rate. One of the most critical classification difficulties, thyroid illness categorization, is suggested in this study [11]. The two types of thyroid glands—hypothyroid and hyperthyroid—are used for metabolic regulation, and they were categorized utilizing feature extraction and preprocessing techniques like GA (Genetic Algorithm). As a classifier, SVM distinguishes between thyroid diseases. The datasets used in this work come from two sources: first, the machine learning repository at the University of California, Irvine; and second, genuine data collected from the Imam Khomeini hospital by the Intelligent System Laboratory at the K.N. Toosi University of Technology. According to this study, early stage heart disease prediction has previously been done, but this paper proposes to use feature selection techniques as a Rapid miner tool to improve prediction accuracy. The algorithms used are Decision Tree, Logistic Regression, Naïve Bayes, and Random Forest, with respective accuracy rates of 82.22%, 82.56%, 84.17%, 84.24%, and 84.85%. Research has been conducted using datasets extracted from the UCI dataset. The authors of [13] put out a model for diagnosing thyroid disease using tools like Naive Bayes, Support Vector Machines (SVMs), and Random Forest as classification algorithms, as well as feature selection methods like Univariate Selection (UFS), Recursive Feature Elimination (RFE), and Tree-Based Feature Selection. They get the best accuracy of 92.92% while using SVM in conjunction with RFE. They retrieved the dataset from the UCI Machine Learning Repository and used it for their study.

III. METHODOLOGY

A common adage in machine learning states that you will only obtain trash value out of a machine learning system if you feed it rubbish value. When using a machine learning algorithm to make a prediction, it becomes more difficult for the algorithms to reach optimal accuracy if the data set comprises irrelevant and noisy data. We want to provide the algorithm the most significant characteristics in order to attain the best accuracy possible, and the feature selection approach helps us accomplish just that. First, we cleaned up the data after collecting hypothyroid information from the certified diagnostic clinic. Step two included using feature selection techniques like RFE, UFS, and PCA on our dataset to identify key properties. In the third stage, we evaluate the algorithms separately by using the features that were selected. We used the following classification techniques to analyze our dataset: Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB). In figure 1 we can see the structure.

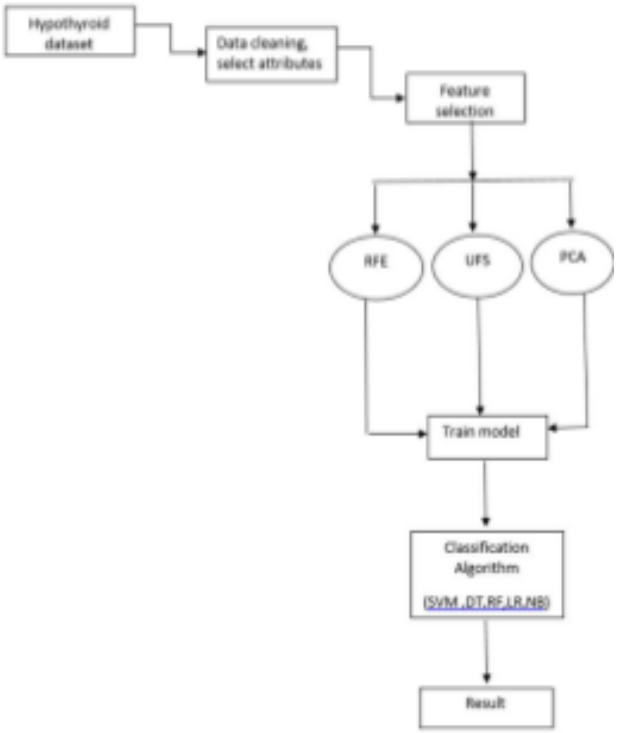

Fig-1: Data flow of the model

Section A. Dataset Overview We had an extremely difficult time collecting data in 2020 due to the pandemic condition. Datasets were retrieved from a Dhaka, Bangladesh, certified diagnostic clinic. We gathered 519 pieces of data with 9 different qualities in total. The table in the dataset has the following characteristics. 1-

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

Attributes	Type	Descripti
ID	Continuous	Patients I
Age	Continuous	In years
Sex	Male , Female	Gender
FT3	Continuous	Free Triiodothyro value
FT4	Continuous	Free Thyro value
T3	Continuous	Triiodothyro value
T4	Continuous	Thyroxin v
TSH	Continuous	Thyroid Stimulati Hormone v
Result	categorical	0/1

Table-1: Attributes of Hypothyroid Dataset

Method B. For Feature Selection Feature selection is a method for automatically picking out the most relevant characteristics to use in making predictions about the outputs or variables of interest. Our model's accuracy is severely compromised by some variables in our dataset. A key function for feature selection approach is the elimination of such unnecessary data. Feature selection maximises the likelihood of reaching a conclusion based on relevant information and reduces overfitting by making data less superfluous. 2. Enhanced Precision — It cleans up our data so it's less deceptive, which in turn improves the precision of our models. Thirdly, less data equals quicker training by lowering both the complexity and the amount of time needed to train the algorithm. C. Feature Selection Method One approach to feature selection is recursive feature elimination (RFE). This technique involves fitting a model and then ranking the features based on the 'coef_' and 'feature_importances_' characteristics. The goal is to remove fragile features. The 'fit' approach determines which traits are most significant, and then it iteratively removes the ones that aren't until it gets to the ones you want. We identified three critical characteristics per algorithm after running RFE in a number of them to determine which ones were most suited to each. The following algorithms have their estimated accuracy using RFE: SVM(99.35%), Decision Tree(99.35%), Random Forest(99.35%), Logistic Regression(99.35%), and Naive Bayes (94.23%).

Feature Selection Technique	Algorithm	Importance feature
RFE	SVM	T3,T4,TSH
RFE	Decision Tree	T3,T4,TSH
RFE	Random Forest	Age,FT4,TSH
RFE	Logistic Regression	FT3,T3,TSH
RFE	Naïve Bayes	T3,T4,TSH

Table-2: RFE Feature Selection

C2. Univariate Feature Selection (UFS): UFS is an additional feature selection technique that identifies the features with the highest scores by combining the 'SelectKBest' with a chi-squared test (score_func=chi2). The chi-squared test is a statistical method that measures the strength of the link between features and the response variable. From our dataset, our approach extracts three crucial properties. Each algorithm's predicted accuracy using UFS is as follows: SVM (98.71%), Decision Tree (99.35%), Random Forest (99.35%), Logistic Regression (99.35%), and Naive Bayes (96.79%).

Feature Selection Technique	Algorithm	Importance feature
UFS	SVM , Decision Tree, Random Forest, Logistic Regression, Naïve Bayes	Age,T4,TSH

Table-3: UFS Feature Selection

C3. Principal Component Analysis (PCA): PCA is a data reduction approach that is a crucial feature selection tool. It takes high-dimensional data and makes it low-dimensional so that the most relevant features may capture all the information in the dataset. The 'explained_variance_ratio_' property ranks important features; the feature responsible for the largest PCA variance is considered the first principal component, the feature responsible for the second largest variance is considered the second principal component, and so on. The following algorithms were found to have an estimated accuracy of 89.74% when employing PCA: SVM, Decision Tree, Random Forest, Logistic Regression, and Naive Bayes.

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

Feature Selection Technique	Algorithm	Importance featur
PCA	SVM , Decision Tree, Random Forest, Logistic Regression, Naïve Bayes	Age,FT3,FT4

Table-4: PCA Feature Selection Algorithm

IV. RESULT ANALYSIS

In our model, we used three feature selection strategies to forecast hypothyroidism. Additionally, it demonstrated that early hypothyroidism may be predicted using a machine learning method. To improve the algorithm's performance and determine the optimal feature selection strategy for our model, we used RFE, UFS, and PCA feature selection. We looked for significant attributes and identified them. Table 5 shows that algorithms benefit from the RFE feature selection approach when it comes to choosing appropriate characteristics. Hence, the RFE feature selection method outperforms the other two approaches, maintaining an accuracy level of 99.35% throughout all four methods. In contrast, principal component analysis (PCA) yields the most inaccurate results of any of these techniques.

Serial Number	Algorithm	Feature Selection Technique (RFE) Accuracy%	Feature Selection Technique (UFS) Accuracy%	Feature Selection Technique (PCA) Accuracy%
1	SVM	99.35%	98.71%	89.74%
2	Decision Tree	99.35%	99.35%	87.17%
3	Random Forest	99.35%	99.35%	88.46%
4	Logistic Regression	99.35%	99.35%	89.74%
5	Naïve Bayes	94.23%	96.79%	89.74%

Table-5: Result Analysis

V. CONCLUSION

Compared to the other classifiers, RFE, a feature selection approach, significantly improves our accuracy. We found that RFE greatly improves our ability to use a real-time dataset to forecast primary-stage hypothyroidism. Given the present pandemic scenario, data collection is proving to be rather challenging for us. Therefore, we have only managed to gather 519 pieces of data. Therefore, due to the circumstances and the limitation, we were unable to conduct the investigation on a bigger dataset. Based on our research, it seems that no prior studies have focused on thyroid in Bangladesh. The amount of data we have at our disposal is limited. Consequently, we want to expand our dataset in the future and encourage more individuals from our nation to get involved in this condition so that we may discover a better solution and make more accurate primary stage disease predictions. I pray that it will aid our nation's citizens in keeping society in good shape.

REFERENCES

[1] A. M. Amiri, and G. Armano, “Early Diagnosis of Heart Disease Using Classification And Regression Trees”, In The 2013 International Joint Conference on Neural Networks, pp. 1-4, 09 January, 2014.

[2] A. K. Aswathi, and A. Antony, “An Intelligent System for Thyroid Disease Classification and Diagnosis”, 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018), pp. 1261-1264, 27 September, 2018.

[3] A. Begum, and A. Parkavi, “Prediction of thyroid Disease Using Data Mining Techniques”, 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 342-345, 06 June, 2019.

[4] K. Pavya, and B. Srinivasan, “FEATURE SELECTION ALGORITHMS TO IMPROVE THYROID DISEASE DIAGNOSIS”, IEEE International Conference on Innovations in Green Energy and Healthcare Technologies(ICIGEHT’17), pp. 1-5, 02 November, 2017.

[5] F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab, “Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM”, 3rd

International Conference on Bioinformatics and Biomedical Engineering, pp. 1-4, 14 July, 2009.

[6] Q. Pan, , Y. Zhang, M. Zuo, L. Xiang, and D. Chen, "Improved Ensemble Classification Method of Thyroid Disease Based on Random Forest", 8th International Conference on Information Technology in Medicine and Education, pp 567-571, 13 July, 2017.

[7] A. Tyagi, R. Mehra, and A. Saxena, "Interactive Thyroid Disease Prediction System Using Machine Learning Technique", 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), pp 689-693, 27 June, 2019.

[8] S. Dash, M. N. Das, and B. K. Mishra, "Implementation of an Optimized Classification Model for Prediction of Hypothyroid Disease Risks", International Conference on Inventive Computation Technologies (ICICT) ,pp. 1-4, 19 January, 2017.

[9] V. S. Vairale, and S. Shukla, "Classification of Hypothyroid Disorder using Optimized SVM Method", Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019), pp. 258-263, 10 February, 2020.

[10] K. Shankar, S. K. Lakshmanaprabu, D. Gupta, A. Maseleno, V. H. C. D. Albuquerque, "Optimal feature-based multi-kernel SVM approach for thyroid disease classification", Springer Science +Business Media, LLC, part of Springer Nature 2018, pp. 1128-1143, 2 July, 2018.

[11] M. R. N. Kousarrizi, F.Seiti, and M. Teshnehlab, "An Experimental Comparative Study on Thyroid Disease Diagnosis Based on Feature Subset Selection and classification", International Journal of Electrical & Computer Sciences IJECS-IJENS, pp. 13-19, February, 2012.

[12] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches", 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST), pp. 619-623, 18 March, 2019.

[13] P. Duggal, and S. Shukla, "Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques", 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 670-675, 09 April 2020.

[14] Dhaka Tribune(2018), 50 million people suffer from thyroid disease in Bangladesh. Available: https://www.dhakatribune.com/feature/healthwellness/2018/05/25/experts-50-million-people-suffer-from-thyroiddisease-in-bangladesh/.